

ISCA 2016

2016 43rd International Symposium on Computer Architecture

18-22 June 2016 · Seoul, South Korea

[CONFERENCE INFORMATION](#)

[PAPERS BY SESSION](#)

[PAPERS BY AUTHOR](#)

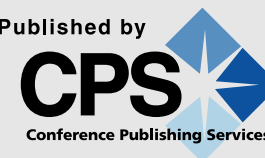
[GETTING STARTED](#)

[TRADEMARKS](#)

[SEARCH](#)



Published by



IEEE  computer society
CELEBRATING 70 YEARS

Conference Information

2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture

- ☐ Message from the General Chairs
- ☐ Message from the Program Chair
- ☐ Organizing Committee
- ☐ Program Committee
- ☐ External Reviewers
- ☐ Sponsors
- ☐ Title Page (Book version)
- ☐ Copyright Page (Book version)
- ☐ Table of Contents (Book version)
- ☐ Author Index (Book version)
- ☐ Publisher's Information (Book version)

Sessions

- ☐ Session 1A: Neural Networks I
- ☐ Session 1B: Heterogeneous Architecture/ Approximate Computing
- ☐ Session 2A: Caches
- ☐ Session 2B: Hardware Design
- ☐ Session 3A: Accelerators
- ☐ Session 3B: GPU I
- ☐ Session 4A: Neural Networks II
- ☐ Session 4B: NoC/Virtualization
- ☐ Session 5A: Cache/Memory Compression
- ☐ Session 5B: Reliability I
- ☐ Session 6: Neural Networks III
- ☐ Session 7A: Micro Architecture
- ☐ Session 7B: Datacenter
- ☐ Session 8A: Memory I
- ☐ Session 8B: Emerging Architectures
- ☐ Session 9A: GPU II

Sessions (cont.)

- ☐ Session 9B: Reliability II
- ☐ Session 10A: Energy Efficient Computing
- ☐ Session 10B: Memory II

Papers by Session

Session 1A: Neural Networks I

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing
Jorge Albericio, Patrick Judd, Tayler Hetherington, Tor Aamodt, Natalie Enright Jerger, and Andreas Moshovos
- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars
Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R. Stanley Williams, and Vivek Srikumar
- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory
Ping Chi, Shuangchen Li, Cong Xu, Tao Zhang, Jishen Zhao, Yongpan Liu, Yu Wang, and Yuan Xie

Papers by Session

Session 1B: Heterogeneous Architecture/ Approximate Computing

- ❑ Asymmetry-Aware Work-Stealing Runtimes
Christopher Torng, Moyang Wang, and Christopher Batten
- ❑ Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing
Hung-Wei Tseng, Qianchen Zhao, Yuxiao Zhou, Mark Gahagan, and Steven Swanson
- ❑ Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration
Divya Mahajan, Amir Yazdanbaksh, Jongse Park, Bradley Thwaites, and Hadi Esmaeilzadeh

Papers by Session

Session 2A: Caches

- ❑ Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement
Akanksha Jain and Calvin Lin
- ❑ Efficient Synonym Filtering and Scalable Delayed Translation for Hybrid Virtual Caching
Chang Hyun Park, Taekyung Heo, and Jaehyuk Huh
- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches
Hsiang-Yun Cheng, Jishen Zhao, Jack Sampson, Mary Jane Irwin, Aamer Jaleel, Yu Lu, and Yuan Xie

Papers by Session

Session 2B: Hardware Design

- ❑ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware
David Koeplinger, Raghu Prabhakar, Yaqi Zhang, Christina Delimitrou, Christos Kozyrakis, and Kunle Olukotun
- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL
Donggyu Kim, Adam Izraelevitz, Christopher Celio, Hokeun Kim, Brian Zimmer, Yunsup Lee, Jonathan Bachrach, and Krste Asanovic
- ❑ PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures
Michael A. Laurenzano, Yunqi Zhang, Jiang Chen, Lingjia Tang, and Jason Mars

Papers by Session

Session 3A: Accelerators

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads
Boncheol Gu, Andre S. Yoon, Duck-Ho Bae, Insoon Jo, Jinyoung Lee, Jonghyun Yoon, Jeong-Uk Kang, Moonsang Kwon, Chanhoo Yoon, Sangyeun Cho, Jaeheon Jeong, and Duckhyun Chang
- ❑ Energy Efficient Architecture for Graph Analytics Accelerators
Muhammet Mustafa Ozdal, Serif Yesil, Taemin Kim, Andrey Ayupov, John Greth, Steven Burns, and Ozcan Ozturk
- ❑ ASIC Clouds: Specializing the Datacenter
Ikuo Magaki, Moein Khazraee, Luis Vega Gutierrez, and Michael Bedford Taylor

Papers by Session

Session 3B: GPU I

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs
Yunho Oh, Keunsoo Kim, Myung Kuk Yoon, Jong Hyun Park, Yongjun Park, Won Woo Ro, and Murali Annavaram
- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems
Kevin Hsieh, Eiman Ebrahim, Gwangsun Kim, Niladrish Chatterjee, Mike O'Connor, Nandita Vijaykumar, Onur Mutlu, and Stephen W. Keckler
- ❑ Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming
Chang Hyun Park, Taekyung Heo, and Jaehyuk Huh
- ❑ Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming
Qiumin Xu, Hyeran Jeon, Keunsoo Kim, Won Woo Ro, and Murali Annavaram

Papers by Session

Session 4A: Neural Networks II

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network
Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally
- ❑ RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision
Robert LiKamWa, Yunhui Hou, Yuan Gao, Mia Polansky, and Lin Zhong
- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators
Brandon Reagen, Paul Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee, Sae Kyu Lee, José Miguel Hernández-Lobato, Gu-Yeon Wei, and David Brooks

Papers by Session

Session 4B: NoC/Virtualization

- ❑ Opportunistic Competition Overhead Reduction for Expediting Critical Section in NoC Based CMPs
Yuan Yao and Zhonghai Lu
- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors
Channoh Kim, Sungmin Kim, Hyeon Gyu Cho, Dooyoung Kim, Jaehyeok Kim, Young H. Oh, Hakbeom Jang, and Jae W. Lee
- ❑ ARM Virtualization: Performance and Architectural Implications
Christoffer Dall, Shih-Wei Li, Jin Tack Lim, Jason Nieh, and Georgios Koloventzos

Papers by Session

Session 5A: Cache/Memory Compression

- ❑ Base-Victim Compression: An Opportunistic Cache Compression Architecture
Jayesh Gaur, Alaa R. Alameldeen, and Sreenivas Subramoney
- ❑ Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures
Jungrae Kim, Michael Sullivan, Esha Choukse, and Mattan Erez

Papers by Session

Session 5B: Reliability I

- ❑ XED: Exposing On-Die Error Detection Information for Strong Memory Reliability
Prashant J. Nair, Vilas Sridharan, and Moinuddin K. Qureshi
- ❑ Production-Run Software Failure Diagnosis via Adaptive Communication Tracking
Mohammad Mejbah Ul Alam and Abdullah Muzahid

Papers by Session

Session 6: Neural Networks III

- ❑ Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks
Yu-Hsin Chen, Joel Emer, and Vivienne Sze
- ❑ Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory
Duckhwan Kim, Jaeha Kung, Sek Chai, Sudhakar Yalamanchili, and Saibal Mukhopadhyay
- ❑ Cambricon: An Instruction Set Architecture for Neural Networks
Shaoli Liu, Zidong Du, Jinhua Tao, Dong Han, Tao Luo, Yuan Xie, Yunji Chen, and Tianshi Chen

Papers by Session

Session 7A: Micro Architecture

- ❑ Decoupling Loads for Nano-Instruction Set Computers
Ziqiang Huang, Andrew D. Hilton, and Benjamin C. Lee
- ❑ Future Vector Microprocessor Extensions for Data Aggregations
Timothy Hayes, Oscar Palomar, Osman Unsal, Adrian Cristal, and Mateo Valero
- ❑ Efficiently Scaling Out-of-Order Cores for Simultaneous Multithreading
Faissal M. Sleiman and Thomas F. Wenisch
- ❑ Accelerating Dependent Cache Misses with an Enhanced Memory Controller
Milad Hashemi, Khubaib, Eiman Ebrahimi, Onur Mutlu, and Yale N. Patt

Papers by Session

Session 7B: Datacenter

- ❑ Treadmill: Attributing the Source of Tail Latency through Precise Load Testing and Statistical Inference
Yunqi Zhang, David Meisner, Jason Mars, and Lingjia Tang
- ❑ Dynamo: Facebook's Data Center-Wide Power Management System
Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song
- ❑ Peak Efficiency Aware Scheduling for Highly Energy Proportional Servers
Daniel Wong
- ❑ Power Attack Defense: Securing Battery-Backed Data Centers
Chao Li, Zhenhua Wang, Xiaofeng Hou, Haopeng Chen, Xiaoyao Liang, and Minyi Guo

Papers by Session

Session 8A: Memory I

- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric
Mingyu Gao, Christina Delimitrou, Dimin Niu, Krishna T. Malladi, Hongzhong Zheng, Bob Brennan, and Christos Kozyrakis
- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs
Lunkai Zhang, Brian Neely, Diana Franklin, Dmitri Strukov, Yuan Xie, and Frederic T. Chong
- ❑ MITTS: Memory Inter-arrival Time Traffic Shaping
Yanqi Zhou and David Wentzlaff

Papers by Session

Session 8B: Emerging Architectures

- ❑ The Anytime Automaton
Joshua San Miguel and Natalie Enright Jerger
- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units
Siyang Wang, Xiangyu Zhang, Yuxuan Li, Ramin Bashizade, Song Yang, Chris Dwyer, and Alvin R. Lebeck
- ❑ Evaluation of an Analog Accelerator for Linear Algebra
Yipeng Huang, Ning Guo, Mingoo Seok, Yannis Tsividis, and Simha Sethumadhavan

Papers by Session

Session 9A: GPU II

- ❑ LaPerm: Locality Aware Scheduler for Dynamic Parallelism on GPUs
Jin Wang, Norm Rubin, Albert Sidelnik, and Sudhakar Yalamanchili
- ❑ ActivePointers: A Case for Software Address Translation on GPUs
Sagi Shahar, Shai Bergman, and Mark Silberstein
- ❑ Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit
Myung Kuk Yoon, Keunsoo Kim, Sangpil Lee, Won Woo Ro, and Murali Annavaram

Papers by Session

Session 9B: Reliability II

- ❑ All-Inclusive ECC: Thorough End-to-End Protection for Reliable Computer Memory
Jungrae Kim, Michael Sullivan, Sangkug Lym, and Mattan Erez
- ❑ Rescuing Uncorrectable Fault Patterns in On-Chip Memories through Error Pattern Transformation
Henry Duwe, Xun Jian, Daniel Petrisko, and Rakesh Kumar
- ❑ RelaxFault Memory Repair
Dong Wan Kim and Mattan Erez

Papers by Session

Session 10A: Energy Efficient Computing

- ❑ Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures
Raghavendra Pradyumna Pothukuchi, Amin Ansari, Petros Voulgaris, and Josep Torrellas
- ❑ Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems
Hari Cherupalli, Rakesh Kumar, and John Sartori
- ❑ CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture
Yanqi Zhou, Henry Hoffmann, and David Wentzlaff

Papers by Session

Session 10B: Memory II

- ❑ Boosting Access Parallelism to PCM-Based Main Memory
Mohammad Arjomand, Mahmut T. Kandemir, Anand Sivasubramaniam, and Chita R. Das
- ❑ Agile Paging: Exceeding the Best of Nested and Shadow Paging
Jayneel Gandhi, Mark D. Hill, and Michael M. Swift
- ❑ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity
Hoseok Seol, Wongyu Shin, Jaemin Jang, Jungwhan Choi, Jinwoong Suh, and Lee-Sup Kim

Papers by Author

A

- ☐ Aamodt, Tor
- ☐ Adolf, Robert
- ☐ Alam, Mohammad Mejbah UI
- ☐ Alameldeen, Alaa R.
- ☐ Albericio, Jorge
- ☐ Annavaram, Murali
- ☐ Ansari, Amin
- ☐ Arjomand, Mohammad
- ☐ Asanovicc, Krste
- ☐ Ayupov, Andrey

B

- ☐ Bachrach, Jonathan
- ☐ Bae, Duck-Ho
- ☐ Balasubramonian, Rajeev

- ☐ Bashizade, Ramin
- ☐ Batten, Christopher
- ☐ Bergman, Shai
- ☐ Brennan, Bob
- ☐ Brooks, David
- ☐ Burns, Steven

C

- ☐ Celio, Christopher
- ☐ Chai, Sek
- ☐ Chang, Duckhyun
- ☐ Chatterjee, Niladrish
- ☐ Chen, Haopeng
- ☐ Chen, Jiang
- ☐ Chen, Tianshi
- ☐ Chen, Yu-Hsin
- ☐ Chen, Yunji

Papers by Author

- ☐ Cheng, Hsiang-Yun
- ☐ Cherupalli, Hari
- ☐ Chi, Ping
- ☐ Cho, Hyeon Gyu
- ☐ Cho, Sangyeun
- ☐ Choi, Jungwhan
- ☐ Chong, Frederic T.
- ☐ Choukse, Esha
- ☐ Cristal, Adrian

D

- ☐ Dall, Christoffer
- ☐ Dally, William J.
- ☐ Das, Chita R.
- ☐ Delimitrou, Christina
- ☐ Deng, Qingyuan
- ☐ Du, Zidong

- ☐ Duwe, Henry
- ☐ Dwyer, Chris

E

- ☐ Ebrahim, Eiman
- ☐ Ebrahimi, Eiman
- ☐ Emer, Joel
- ☐ Erez, Mattan
- ☐ Esmaeilzadeh, Hadi

F

- ☐ Franklin, Diana

G

- ☐ Gahagan, Mark
- ☐ Gandhi, Jayneel
- ☐ Ganesh, Lakshmi

Papers by Author

- ☐ Gao, Mingyu
- ☐ Gao, Yuan
- ☐ Gaur, Jayesh
- ☐ Greth, John
- ☐ Gu, Boncheol
- ☐ Guo, Minyi
- ☐ Guo, Ning
- ☐ Gutierrez, Luis Vega

H

- ☐ Han, Dong
- ☐ Han, Song
- ☐ Hashemi, Milad
- ☐ Hayes, Timothy
- ☐ Heo, Taekyung
- ☐ Hernández-Lobato, José Miguel
- ☐ Hetherington, Tayler

- ☐ Hill, Mark D.
- ☐ Hilton, Andrew D.
- ☐ Hoffmann, Henry
- ☐ Horowitz, Mark A.
- ☐ Hou, Xiaofeng
- ☐ Hou, Yunhui
- ☐ Hsieh, Kevin
- ☐ Hsu, Chang-Hong
- ☐ Hu, Miao
- ☐ Huang, Yipeng
- ☐ Huang, Ziqiang
- ☐ Huh, Jaehyuk

I

- ☐ Irwin, Mary Jane
- ☐ Izraelevitz, Adam

Papers by Author

J

- ☐ Jain, Akanksha
- ☐ Jaleel, Aamer
- ☐ Jang, Hakbeom
- ☐ Jang, Jaemin
- ☐ Jeon, Hyeran
- ☐ Jeong, Jaeheon
- ☐ Jerger, Natalie Enright
- ☐ Jian, Xun
- ☐ Jin, Yun
- ☐ Jo, Insoon
- ☐ Judd, Patrick

K

- ☐ Kandemir, Mahmut T.
- ☐ Kang, Jeong-Uk

- ☐ Keckler, Stephen W.
- ☐ Khazraee, Moein
- ☐ Khubaib
- ☐ Kim, Channoh
- ☐ Kim, Dong Wan
- ☐ Kim, Donggyu
- ☐ Kim, Dooyoung
- ☐ Kim, Duckhwan
- ☐ Kim, Gwangsun
- ☐ Kim, Hokeun
- ☐ Kim, Jaehyeok
- ☐ Kim, Jungrae
- ☐ Kim, Keunsoo
- ☐ Kim, Lee-Sup
- ☐ Kim, Sungmin
- ☐ Kim, Taemin

Papers by Author

- ☐ Koeplinger, David
- ☐ Koloventzos, Georgios
- ☐ Kozyrakis, Christos
- ☐ Kumar, Rakesh
- ☐ Kumar, Sanjeev
- ☐ Kung, Jaeha
- ☐ Kwon, Moonsang

L

- ☐ Laurenzano, Michael A.
- ☐ Lebeck, Alvin R.
- ☐ Lee, Benjamin C.
- ☐ Lee, Hyunkwang
- ☐ Lee, Jae W.
- ☐ Lee, Jinyoung
- ☐ Lee, Sae Kyu
- ☐ Lee, Sangpil
- ☐ Lee, Yunsup
- ☐ Li, Bin
- ☐ Li, Chao
- ☐ Li, Shih-Wei
- ☐ Li, Shuangchen
- ☐ Li, Yuxuan
- ☐ Liang, Xiaoyao
- ☐ LiKamWa, Robert
- ☐ Lim, Jin Tack
- ☐ Lin, Calvin
- ☐ Liu, Shaoli
- ☐ Liu, Xingyu
- ☐ Liu, Yongpan
- ☐ Lu, Yu
- ☐ Lu, Zhonghai
- ☐ Luo, Tao

Papers by Author

☐ Lym, Sangkug

M

- ☐ Magaki, Ikuo
- ☐ Mahajan, Divya
- ☐ Malladi, Krishna T.
- ☐ Mao, Huizi
- ☐ Mars, Jason
- ☐ Meisner, David
- ☐ Meza, Justin
- ☐ Miguel, Joshua San
- ☐ Moshovos, Andreas
- ☐ Mukhopadhyay, Saibal
- ☐ Muralimanohar, Naveen
- ☐ Mutlu, Onur
- ☐ Muzahid, Abdullah

N

- ☐ Nag, Anirban
- ☐ Nair, Prashant J.
- ☐ Neely, Brian
- ☐ Nieh, Jason
- ☐ Niu, Dimin

O

- ☐ O'Connor, Mike
- ☐ Oh, Young H.
- ☐ Oh, Yunho
- ☐ Olukotun, Kunle
- ☐ Ozdal, Muhammet Mustafa
- ☐ Ozturk, Ozcan

Papers by Author

P

- ☐ Palomar, Oscar
- ☐ Park, Chang Hyun
- ☐ Park, Jong Hyun
- ☐ Park, Jongse
- ☐ Park, Yongjun
- ☐ Patt, Yale N.
- ☐ Pedram, Ardavan
- ☐ Petrisko, Daniel
- ☐ Polansky, Mia
- ☐ Pothukuchi, Raghavendra Pradyumna
- ☐ Prabhakar, Raghu
- ☐ Pu, Jing

Q

- ☐ Qureshi, Moinuddin K.

R

- ☐ Rama, Saketh
- ☐ Reagen, Brandon
- ☐ Ro, Won Woo
- ☐ Rubin, Norm

S

- ☐ Sampson, Jack
- ☐ Sartori, John
- ☐ Seok, Mingoo
- ☐ Seol, Hoseok
- ☐ Sethumadhavan, Simha
- ☐ Shafiee, Ali
- ☐ Shahrar, Sagi
- ☐ Shin, Wongyu
- ☐ Sidelnik, Albert

Papers by Author

- ☐ Silberstein, Mark
- ☐ Sivasubramaniam, Anand
- ☐ Sleiman, Faissal M.
- ☐ Song, Yee Jiun
- ☐ Sridharan, Vilas
- ☐ Srikumar, Vivek
- ☐ Strachan, John Paul
- ☐ Strukov, Dmitri
- ☐ Subramoney, Sreenivas
- ☐ Suh, Jinwoong
- ☐ Sullivan, Michael
- ☐ Swanson, Steven
- ☐ Swift, Michael M.
- ☐ Sze, Vivienne

T

- ☐ Tang, Lingjia

- ☐ Tao, Jinhua
- ☐ Taylor, Michael Bedford
- ☐ Thwaites, Bradley
- ☐ Torng, Christopher
- ☐ Torrellas, Josep
- ☐ Tseng, Hung-Wei
- ☐ Tsvidis, Yannis

U

- ☐ Unsal, Osman

V

- ☐ Valero, Mateo
- ☐ Vijaykumar, Nandita
- ☐ Voulgaris, Petros

Papers by Author

W

- ☐ Wang, Jin
- ☐ Wang, Moyang
- ☐ Wang, Siyang
- ☐ Wang, Yu
- ☐ Wang, Zhenhua
- ☐ Wei, Gu-Yeon
- ☐ Wenisch, Thomas F.
- ☐ Wentzlaff, David
- ☐ Whatmough, Paul
- ☐ Williams, R. Stanley
- ☐ Wong, Daniel
- ☐ Wu, Qiang

X

- ☐ Xie, Yuan

- ☐ Xu, Cong
- ☐ Xu, Qiumin

Y

- ☐ Yalamanchili, Sudhakar
- ☐ Yang, Song
- ☐ Yao, Yuan
- ☐ Yazdanbaksh, Amir
- ☐ Yesil, Serif
- ☐ Yoon, Andre S.
- ☐ Yoon, Chanhoo
- ☐ Yoon, Jonghyun
- ☐ Yoon, Myung Kuk

Z

- ☐ Zhang, Lunkai
- ☐ Zhang, Tao

Papers by Author

- ☐ Zhang, Xiangyu
- ☐ Zhang, Yaqi
- ☐ Zhang, Yunqi
- ☐ Zhao, Jishen
- ☐ Zhao, Qianchen
- ☐ Zheng, Hongzhong
- ☐ Zhong, Lin
- ☐ Zhou, Yanqi
- ☐ Zhou, Yuxiao
- ☐ Zimmer, Brian

Papers by Author

Aamodt, Tor

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing

Adolf, Robert

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Alam, Mohammad Mejbah UI

- ❑ Production-Run Software Failure Diagnosis via Adaptive Communication Tracking

Alameldeen, Alaa R.

- ❑ Base-Victim Compression: An Opportunistic Cache Compression Architecture

Albericio, Jorge

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing

Papers by Author

Annavaaram, Murali

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs
- ❑ Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming
- ❑ Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit

Ansari, Amin

- ❑ Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures

Arjomand, Mohammad

- ❑ Boosting Access Parallelism to PCM-Based Main Memory

Asanovicc, Krste

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Papers by Author

Ayupov, Andrey

- ❑ Energy Efficient Architecture for Graph Analytics Accelerators

Bachrach, Jonathan

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Bae, Duck-Ho

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Balasubramonian, Rajeev

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Bashizade, Ramin

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Papers by Author

Batten, Christopher

- ❑ Asymmetry-Aware Work-Stealing Runtimes

Bergman, Shai

- ❑ ActivePointers: A Case for Software Address Translation on GPUs

Brennan, Bob

- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

Brooks, David

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Burns, Steven

- ❑ Energy Efficient Architecture for Graph Analytics Accelerators

Papers by Author

Celio, Christopher

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Chai, Sek

- ❑ Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory

Chang, Duckhyun

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Chatterjee, Niladrish

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Chen, Haopeng

- ❑ Power Attack Defense: Securing Battery-Backed Data Centers

Papers by Author

Chen, Jiang

- ❑ PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures

Chen, Tianshi

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Chen, Yu-Hsin

- ❑ Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks

Chen, Yunji

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Papers by Author

Cheng, Hsiang-Yun

- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches

Cherupalli, Hari

- ❑ Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems

Chi, Ping

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory

Cho, Hyeon Gyu

- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Papers by Author

Cho, Sangyeun

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Choi, Jungwhan

- ❑ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity

Chong, Frederic T.

- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs

Choukse, Esha

- ❑ Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures

Cristal, Adrian

- ❑ Future Vector Microprocessor Extensions for Data Aggregations

Papers by Author

Dall, Christoffer

- ❑ ARM Virtualization: Performance and Architectural Implications

Dally, William J.

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Das, Chita R.

- ❑ Boosting Access Parallelism to PCM-Based Main Memory

Delimitrou, Christina

- ❑ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware
- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

Deng, Qingyuan

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Papers by Author

Du, Zidong

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Duwe, Henry

- ❑ Rescuing Uncorrectable Fault Patterns in On-Chip Memories through Error Pattern Transformation

Dwyer, Chris

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Ebrahim, Eiman

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Ebrahimi, Eiman

- ❑ Accelerating Dependent Cache Misses with an Enhanced Memory Controller

Papers by Author

Emer, Joel

- ❑ Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks

Erez, Mattan

- ❑ Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures
- ❑ All-Inclusive ECC: Thorough End-to-End Protection for Reliable Computer Memory
- ❑ RelaxFault Memory Repair

Esmailzadeh, Hadi

- ❑ Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration

Papers by Author

Franklin, Diana

- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs

Gahagan, Mark

- ❑ Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing

Gandhi, Jayneel

- ❑ Agile Paging: Exceeding the Best of Nested and Shadow Paging

Ganesh, Lakshmi

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Gao, Mingyu

- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

Papers by Author

Gao, Yuan

- ❑ RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision

Gaur, Jayesh

- ❑ Base-Victim Compression: An Opportunistic Cache Compression Architecture

Greth, John

- ❑ Energy Efficient Architecture for Graph Analytics Accelerators

Gu, Boncheol

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Guo, Minyi

- ❑ Power Attack Defense: Securing Battery-Backed Data Centers

Papers by Author

Guo, Ning

- ❑ Evaluation of an Analog Accelerator for Linear Algebra

Gutierrez, Luis Vega

- ❑ ASIC Clouds: Specializing the Datacenter

Han, Dong

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Han, Song

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Hashemi, Milad

- ❑ Accelerating Dependent Cache Misses with an Enhanced Memory Controller

Papers by Author

Hayes, Timothy

- ❑ Future Vector Microprocessor Extensions for Data Aggregations

Heo, Taekyung

- ❑ Efficient Synonym Filtering and Scalable Delayed Translation for Hybrid Virtual Caching
- ❑ Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming

Hernández-Lobato, José Miguel

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Hetherington, Tayler

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing

Papers by Author

Hill, Mark D.

- ❑ Agile Paging: Exceeding the Best of Nested and Shadow Paging

Hilton, Andrew D.

- ❑ Decoupling Loads for Nano-Instruction Set Computers

Hoffmann, Henry

- ❑ CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

Horowitz, Mark A.

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Hou, Xiaofeng

- ❑ Power Attack Defense: Securing Battery-Backed Data Centers

Papers by Author

Hou, Yunhui

- ❑ RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision

Hsieh, Kevin

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Hsu, Chang-Hong

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Hu, Miao

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Huang, Yipeng

- ❑ Evaluation of an Analog Accelerator for Linear Algebra

Papers by Author

Huang, Ziqiang

- ❑ Decoupling Loads for Nano-Instruction Set Computers

Huh, Jaehyuk

- ❑ Efficient Synonym Filtering and Scalable Delayed Translation for Hybrid Virtual Caching
- ❑ Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming

Irwin, Mary Jane

- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches

Izraelevitz, Adam

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Papers by Author

Jain, Akanksha

- ❑ Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement

Jaleel, Aamer

- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches

Jang, Hakbeom

- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Jang, Jaemin

- ❑ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity

Papers by Author

Jeon, Hyeran

- ❑ Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming

Jeong, Jaeheon

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Jerger, Natalie Enright

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing
- ❑ The Anytime Automaton

Jian, Xun

- ❑ Rescuing Uncorrectable Fault Patterns in On-Chip Memories through Error Pattern Transformation

Papers by Author

Jin, Yun

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Jo, Insoon

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Judd, Patrick

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing

Kandemir, Mahmut T.

- ❑ Boosting Access Parallelism to PCM-Based Main Memory

Kang, Jeong-Uk

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Papers by Author

Keckler, Stephen W.

- ☐ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Khazraee, Moein

- ☐ ASIC Clouds: Specializing the Datacenter

Khubaib

- ☐ Accelerating Dependent Cache Misses with an Enhanced Memory Controller

Kim, Channoh

- ☐ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Kim, Dong Wan

- ☐ RelaxFault Memory Repair

Papers by Author

Kim, Donggyu

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Kim, Dooyoung

- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Kim, Duckhwan

- ❑ Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory

Kim, Gwangsun

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Kim, Hokeun

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Papers by Author

Kim, Jaehyeok

- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Kim, Jungrae

- ❑ Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures
- ❑ All-Inclusive ECC: Thorough End-to-End Protection for Reliable Computer Memory

Kim, Keunsoo

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs
- ❑ Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming
- ❑ Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit

Papers by Author

Kim, Lee-Sup

- ☐ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity

Kim, Sungmin

- ☐ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Kim, Taemin

- ☐ Energy Efficient Architecture for Graph Analytics Accelerators

Koeplinger, David

- ☐ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware

Koloventzos, Georgios

- ☐ ARM Virtualization: Performance and Architectural Implications

Papers by Author

Kozyrakis, Christos

- ❑ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware
- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

Kumar, Rakesh

- ❑ Rescuing Uncorrectable Fault Patterns in On-Chip Memories through Error Pattern Transformation
- ❑ Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems

Kumar, Sanjeev

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Kung, Jaeha

- ❑ Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory

Papers by Author

Kwon, Moonsang

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Laurenzano, Michael A.

- ❑ PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures

Lebeck, Alvin R.

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Lee, Benjamin C.

- ❑ Decoupling Loads for Nano-Instruction Set Computers

Papers by Author

Lee, Hyunkwang

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Lee, Jae W.

- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Lee, Jinyoung

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Lee, Sae Kyu

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Papers by Author

Lee, Sangpil

- ❑ Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit

Lee, Yunsup

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Li, Bin

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Li, Chao

- ❑ Power Attack Defense: Securing Battery-Backed Data Centers

Li, Shih-Wei

- ❑ ARM Virtualization: Performance and Architectural Implications

Papers by Author

Li, Shuangchen

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory

Li, Yuxuan

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Liang, Xiaoyao

- ❑ Power Attack Defense: Securing Battery-Backed Data Centers

LiKamWa, Robert

- ❑ RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision

Lim, Jin Tack

- ❑ ARM Virtualization: Performance and Architectural Implications

Papers by Author

Lin, Calvin

- ❑ Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement

Liu, Shaoli

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Liu, Xingyu

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Liu, Yongpan

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory

Papers by Author

Lu, Yu

- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches

Lu, Zhonghai

- ❑ Opportunistic Competition Overhead Reduction for Expediting Critical Section in NoC Based CMPs

Luo, Tao

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Lym, Sangkug

- ❑ All-Inclusive ECC: Thorough End-to-End Protection for Reliable Computer Memory

Magaki, Ikuo

- ❑ ASIC Clouds: Specializing the Datacenter

Papers by Author

Mahajan, Divya

- ❑ Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration

Malladi, Krishna T.

- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

Mao, Huizi

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Mars, Jason

- ❑ PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures
- ❑ Treadmill: Attributing the Source of Tail Latency through Precise Load Testing and Statistical Inference

Papers by Author

Meisner, David

- ❑ Treadmill: Attributing the Source of Tail Latency through Precise Load Testing and Statistical Inference

Meza, Justin

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Miguel, Joshua San

- ❑ The Anytime Automaton

Moshovos, Andreas

- ❑ Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing

Mukhopadhyay, Saibal

- ❑ Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory

Papers by Author

Muralimanohar, Naveen

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Mutlu, Onur

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems
- ❑ Accelerating Dependent Cache Misses with an Enhanced Memory Controller

Muzahid, Abdullah

- ❑ Production-Run Software Failure Diagnosis via Adaptive Communication Tracking

Nag, Anirban

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Papers by Author

Nair, Prashant J.

- ❑ XED: Exposing On-Die Error Detection Information for Strong Memory Reliability

Neely, Brian

- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs

Nieh, Jason

- ❑ ARM Virtualization: Performance and Architectural Implications

Niu, Dimin

- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

O'Connor, Mike

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Papers by Author

Oh, Young H.

- ❑ Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors

Oh, Yunho

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs

Olukotun, Kunle

- ❑ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware

Ozdal, Muhammet Mustafa

- ❑ Energy Efficient Architecture for Graph Analytics Accelerators

Ozturk, Ozcan

- ❑ Energy Efficient Architecture for Graph Analytics Accelerators

Papers by Author

Palomar, Oscar

- ❑ Future Vector Microprocessor Extensions for Data Aggregations

Park, Chang Hyun

- ❑ Efficient Synonym Filtering and Scalable Delayed Translation for Hybrid Virtual Caching
- ❑ Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming

Park, Jong Hyun

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs

Park, Jongse

- ❑ Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration

Papers by Author

Park, Yongjun

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs

Patt, Yale N.

- ❑ Accelerating Dependent Cache Misses with an Enhanced Memory Controller

Pedram, Ardavan

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Petrisko, Daniel

- ❑ Rescuing Uncorrectable Fault Patterns in On-Chip Memories through Error Pattern Transformation

Polansky, Mia

- ❑ RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision

Papers by Author

Pothukuchi, Raghavendra Pradyumna

- ❑ Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures

Prabhakar, Raghu

- ❑ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware

Pu, Jing

- ❑ EIE: Efficient Inference Engine on Compressed Deep Neural Network

Qureshi, Moinuddin K.

- ❑ XED: Exposing On-Die Error Detection Information for Strong Memory Reliability

Rama, Saketh

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Papers by Author

Reagen, Brandon

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Ro, Won Woo

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs
- ❑ Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming
- ❑ Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit

Rubin, Norm

- ❑ LaPerm: Locality Aware Scheduler for Dynamic Parallelism on GPUs

Sampson, Jack

- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches

Papers by Author

Sartori, John

- ❑ Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems

Seok, Mingoo

- ❑ Evaluation of an Analog Accelerator for Linear Algebra

Seol, Hoseok

- ❑ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity

Sethumadhavan, Simha

- ❑ Evaluation of an Analog Accelerator for Linear Algebra

Papers by Author

Shafiee, Ali

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Shahar, Sagi

- ❑ ActivePointers: A Case for Software Address Translation on GPUs

Shin, Wongyu

- ❑ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity

Sidelnik, Albert

- ❑ LaPerm: Locality Aware Scheduler for Dynamic Parallelism on GPUs

Silberstein, Mark

- ❑ ActivePointers: A Case for Software Address Translation on GPUs

Papers by Author

Sivasubramaniam, Anand

- ❑ Boosting Access Parallelism to PCM-Based Main Memory

Sleiman, Faissal M.

- ❑ Efficiently Scaling Out-of-Order Cores for Simultaneous Multithreading

Song, Yee Jiun

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Sridharan, Vilas

- ❑ XED: Exposing On-Die Error Detection Information for Strong Memory Reliability

Srikumar, Vivek

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Papers by Author

Strachan, John Paul

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Strukov, Dmitri

- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs

Subramoney, Sreenivas

- ❑ Base-Victim Compression: An Opportunistic Cache Compression Architecture

Suh, Jinwoong

- ❑ Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity

Papers by Author

Sullivan, Michael

- ❑ Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures
- ❑ All-Inclusive ECC: Thorough End-to-End Protection for Reliable Computer Memory

Swanson, Steven

- ❑ Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing

Swift, Michael M.

- ❑ Agile Paging: Exceeding the Best of Nested and Shadow Paging

Sze, Vivienne

- ❑ Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks

Papers by Author

Tang, Lingjia

- ❑ PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures
- ❑ Treadmill: Attributing the Source of Tail Latency through Precise Load Testing and Statistical Inference

Tao, Jinhua

- ❑ Cambricon: An Instruction Set Architecture for Neural Networks

Taylor, Michael Bedford

- ❑ ASIC Clouds: Specializing the Datacenter

Thwaites, Bradley

- ❑ Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration

Papers by Author

Torng, Christopher

- ❑ Asymmetry-Aware Work-Stealing Runtimes

Torrellas, Josep

- ❑ Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures

Tseng, Hung-Wei

- ❑ Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing

Tsividis, Yannis

- ❑ Evaluation of an Analog Accelerator for Linear Algebra

Unsal, Osman

- ❑ Future Vector Microprocessor Extensions for Data Aggregations

Papers by Author

Valero, Mateo

- ❑ Future Vector Microprocessor Extensions for Data Aggregations

Vijaykumar, Nandita

- ❑ Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems

Voulgaris, Petros

- ❑ Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures

Wang, Jin

- ❑ LaPerm: Locality Aware Scheduler for Dynamic Parallelism on GPUs

Wang, Moyang

- ❑ Asymmetry-Aware Work-Stealing Runtimes

Papers by Author

Wang, Siyang

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Wang, Yu

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory

Wang, Zhenhua

- ❑ Power Attack Defense: Securing Battery-Backed Data Centers

Wei, Gu-Yeon

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Wenisch, Thomas F.

- ❑ Efficiently Scaling Out-of-Order Cores for Simultaneous Multithreading

Papers by Author

Wentzlaff, David

- ❑ MITTS: Memory Inter-arrival Time Traffic Shaping
- ❑ CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

Whatmough, Paul

- ❑ Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators

Williams, R. Stanley

- ❑ ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars

Wong, Daniel

- ❑ Peak Efficiency Aware Scheduling for Highly Energy Proportional Servers

Papers by Author

Wu, Qiang

- ❑ Dynamo: Facebook's Data Center-Wide Power Management System

Xie, Yuan

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory
- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches
- ❑ Cambricon: An Instruction Set Architecture for Neural Networks
- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs

Xu, Cong

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory

Papers by Author

Xu, Qiumin

- ❑ Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming

Yalamanchili, Sudhakar

- ❑ Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory
- ❑ LaPerm: Locality Aware Scheduler for Dynamic Parallelism on GPUs

Yang, Song

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Yao, Yuan

- ❑ Opportunistic Competition Overhead Reduction for Expediting Critical Section in NoC Based CMPs

Papers by Author

Yazdanbaksh, Amir

- ❑ Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration

Yesil, Serif

- ❑ Energy Efficient Architecture for Graph Analytics Accelerators

Yoon, Andre S.

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Yoon, Chanh

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Yoon, Jonghyun

- ❑ Biscuit: A Framework for Near-Data Processing of Big Data Workloads

Papers by Author

Yoon, Myung Kuk

- ❑ APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs
- ❑ Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit

Zhang, Lunkai

- ❑ Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs

Zhang, Tao

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory

Zhang, Xiangyu

- ❑ Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units

Papers by Author

Zhang, Yaqi

- ❑ Automatic Generation of Efficient Accelerators for Reconfigurable Hardware

Zhang, Yunqi

- ❑ PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures
- ❑ Treadmill: Attributing the Source of Tail Latency through Precise Load Testing and Statistical Inference

Zhao, Jishen

- ❑ PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory
- ❑ LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches

Zhao, Qianchen

- ❑ Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing

Papers by Author

Zheng, Hongzhong

- ❑ DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric

Zhong, Lin

- ❑ RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision

Zhou, Yanqi

- ❑ MITTS: Memory Inter-arrival Time Traffic Shaping
- ❑ CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

Zhou, Yuxiao

- ❑ Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing

Zimmer, Brian

- ❑ Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL

Papers by Author

A	B	C	D	E	F	G
H	I	J	K	L	M	N
O	P	Q	R	S	T	U
V	W	X	Y	Z		